D212 Task 1 v2

# Part I: Research Question

Do identifiable patterns exist in the data related to customer churn?

The goal in answering this question is to decrease costs associated with portfolio maintenance. As the cost to on-board a new customer is already known to be 10 times that of retaining an existing customer, one way to cut costs is by raising retention rates among the existing customer base. Therefore, it would be a valuable undertaking for an organization to have the ability to identify patterns in their customer data set, especially as pertaining to customers likely to churn. Armed with this knowledge, new strategic initiatives could be developed as needed to target the retention of highly valued existing customers.

The research question will be addressed using the k-means clustering technique.

# Part II: Method Justification

The k-means clustering method is chosen as it functions to identify patterns in data. In this case, the research question is indefinite in that it is not aiming to predict a specific response value. Had the research question been aimed at predicting responses of known variables, then supervised machine learning techniques would have been applicable. As the research question is not to predict a response, but to detect patterns in a data set, unsupervised machine learning techniques are more appropriate.

A k-means clustering algorithm functions in iterations to analyze and assign each observation in the data set to a corresponding sub group called a cluster which it shares with other similar data points. The clusters will be non overlapping and therefore homogeneous. After each iteration, the euclidean distance from each observation to its assigned cluster center is calculated. These distances are then squared and summed and the sum of squares for each iteration is compared to return the iteration having the lowest sum of squares. The lowest sum of squares then indicates the most accurate iteration with the observations most closely assigned to their clusters.

The assumption of the k-means technique is that the number of subgroups or clusters is already known. Since in reality the number of clusters is not always known, scree plots become a useful tool in identifying the number of clusters prior to the k-means algorithm.

K means is a useful algorithm for this research question because once plotted we can perform the clustering on a minimal number of variables at a time and begin to see patterns emerge in the clustered data. Since the algorithm sorts the data into clusters for us, the analyst can then analyze the clustered data to find for instance age thresholds or income groupings of telecommunications customers within the data set that might not otherwise have been visible with a large number of observations without clustering.

R will be used for this analysis. R is open source software that was specifically made for statistical analysis. Using R, we can ingest the data set, and leveraging an extensive library of data manipulation and visualization packages, perform the necessary clustering steps. More information can be found on the R project website (<https://www.r-project.org/>).

The dplyr package will be used for data preparation and manipulation within R. More information for the dplyr package can be found on the tidyverse website (<https://dplyr.tidyverse.org/>).

Algorithm construction will be done using the k-means function in base R. The k-means function accepts the data frame and number of clusters as input, along with a parameter determining the number of iterations that the algorithm will go through to find the best fit. No further libraries are necessary for this exercise.

# Part III: Data Preparation

To prepare the data, first a check is run for missing values in the data set using the sapply function. No action is needed as no missing values were detected in the provided data set. However, had missing values been detected more analysis would have been required for each instance to determine handling of the records containing nulls. Depending on the variable in view, the nulls could be filled with variable mean or median values, or even removed from the data set.

Second, prior to creating the k-means model, the data will be prepared by selecting only the numeric variables to be used in model construction. Categorical variables are not appropriate for a k-means model. In variable selection it is necessary to investigate the data set to distinguish between numeric and categorical variables. This is done using R’s str function. From the output, the relevant numeric variables can then be selected.

Once the variables are selected, the data frame is inspected for scaling. If wide value ranges are present, the data variables should be scaled so as to avoid one variable carrying more weight than another.

ready for model construction and the cleaned data is written to a data file.

library(dplyr)

## Warning: replacing previous import 'vctrs::data\_frame' by 'tibble::data\_frame'  
## when loading 'dplyr'

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# Load Data Set  
df<-read.csv("c:/users/shua/documents/Data Mining II\_D212/churn\_clean.csv")  
  
# Check for missing values  
sapply(df, function(x) sum(is.na(x)))

## CaseOrder Customer\_id Interaction   
## 0 0 0   
## UID City State   
## 0 0 0   
## County Zip Lat   
## 0 0 0   
## Lng Population Area   
## 0 0 0   
## TimeZone Job Children   
## 0 0 0   
## Age Income Marital   
## 0 0 0   
## Gender Churn Outage\_sec\_perweek   
## 0 0 0   
## Email Contacts Yearly\_equip\_failure   
## 0 0 0   
## Techie Contract Port\_modem   
## 0 0 0   
## Tablet InternetService Phone   
## 0 0 0   
## Multiple OnlineSecurity OnlineBackup   
## 0 0 0   
## DeviceProtection TechSupport StreamingTV   
## 0 0 0   
## StreamingMovies PaperlessBilling PaymentMethod   
## 0 0 0   
## Tenure MonthlyCharge Bandwidth\_GB\_Year   
## 0 0 0   
## Item1 Item2 Item3   
## 0 0 0   
## Item4 Item5 Item6   
## 0 0 0   
## Item7 Item8   
## 0 0

# Identify numeric columns for use  
str(df)

## 'data.frame': 10000 obs. of 50 variables:  
## $ CaseOrder : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Customer\_id : chr "K409198" "S120509" "K191035" "D90850" ...  
## $ Interaction : chr "aa90260b-4141-4a24-8e36-b04ce1f4f77b" "fb76459f-c047-4a9d-8af9-e0f7d4ac2524" "344d114c-3736-4be5-98f7-c72c281e2d35" "abfa2b40-2d43-4994-b15a-989b8c79e311" ...  
## $ UID : chr "e885b299883d4f9fb18e39c75155d990" "f2de8bef964785f41a2959829830fb8a" "f1784cfa9f6d92ae816197eb175d3c71" "dc8a365077241bb5cd5ccd305136b05e" ...  
## $ City : chr "Point Baker" "West Branch" "Yamhill" "Del Mar" ...  
## $ State : chr "AK" "MI" "OR" "CA" ...  
## $ County : chr "Prince of Wales-Hyder" "Ogemaw" "Yamhill" "San Diego" ...  
## $ Zip : int 99927 48661 97148 92014 77461 31030 37847 73109 34771 45237 ...  
## $ Lat : num 56.3 44.3 45.4 33 29.4 ...  
## $ Lng : num -133.4 -84.2 -123.2 -117.2 -95.8 ...  
## $ Population : int 38 10446 3735 13863 11352 17701 2535 23144 17351 20193 ...  
## $ Area : chr "Urban" "Urban" "Urban" "Suburban" ...  
## $ TimeZone : chr "America/Sitka" "America/Detroit" "America/Los\_Angeles" "America/Los\_Angeles" ...  
## $ Job : chr "Environmental health practitioner" "Programmer, multimedia" "Chief Financial Officer" "Solicitor" ...  
## $ Children : int 0 1 4 1 0 3 0 2 2 1 ...  
## $ Age : int 68 27 50 48 83 83 79 30 49 86 ...  
## $ Income : num 28562 21705 9610 18925 40074 ...  
## $ Marital : chr "Widowed" "Married" "Widowed" "Married" ...  
## $ Gender : chr "Male" "Female" "Female" "Male" ...  
## $ Churn : chr "No" "Yes" "No" "No" ...  
## $ Outage\_sec\_perweek : num 7.98 11.7 10.75 14.91 8.15 ...  
## $ Email : int 10 12 9 15 16 15 10 16 20 18 ...  
## $ Contacts : int 0 0 0 2 2 3 0 0 2 1 ...  
## $ Yearly\_equip\_failure: int 1 1 1 0 1 1 1 0 3 0 ...  
## $ Techie : chr "No" "Yes" "Yes" "Yes" ...  
## $ Contract : chr "One year" "Month-to-month" "Two Year" "Two Year" ...  
## $ Port\_modem : chr "Yes" "No" "Yes" "No" ...  
## $ Tablet : chr "Yes" "Yes" "No" "No" ...  
## $ InternetService : chr "Fiber Optic" "Fiber Optic" "DSL" "DSL" ...  
## $ Phone : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ Multiple : chr "No" "Yes" "Yes" "No" ...  
## $ OnlineSecurity : chr "Yes" "Yes" "No" "Yes" ...  
## $ OnlineBackup : chr "Yes" "No" "No" "No" ...  
## $ DeviceProtection : chr "No" "No" "No" "No" ...  
## $ TechSupport : chr "No" "No" "No" "No" ...  
## $ StreamingTV : chr "No" "Yes" "No" "Yes" ...  
## $ StreamingMovies : chr "Yes" "Yes" "Yes" "No" ...  
## $ PaperlessBilling : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ PaymentMethod : chr "Credit Card (automatic)" "Bank Transfer(automatic)" "Credit Card (automatic)" "Mailed Check" ...  
## $ Tenure : num 6.8 1.16 15.75 17.09 1.67 ...  
## $ MonthlyCharge : num 172 243 160 120 150 ...  
## $ Bandwidth\_GB\_Year : num 905 801 2055 2165 271 ...  
## $ Item1 : int 5 3 4 4 4 3 6 2 5 2 ...  
## $ Item2 : int 5 4 4 4 4 3 5 2 4 2 ...  
## $ Item3 : int 5 3 2 4 4 3 6 2 4 2 ...  
## $ Item4 : int 3 3 4 2 3 2 4 5 3 2 ...  
## $ Item5 : int 4 4 4 5 4 4 1 2 4 5 ...  
## $ Item6 : int 4 3 3 4 4 3 5 3 3 2 ...  
## $ Item7 : int 3 4 3 3 4 3 5 4 4 3 ...  
## $ Item8 : int 4 4 3 3 5 3 5 5 4 3 ...

#Select columns  
df<-df%>%select(Population, Children, Age, Income, Outage\_sec\_perweek, Tenure, MonthlyCharge, Bandwidth\_GB\_Year, Churn)  
  
summary(df)

## Population Children Age Income   
## Min. : 0 Min. : 0.000 Min. :18.00 Min. : 348.7   
## 1st Qu.: 738 1st Qu.: 0.000 1st Qu.:35.00 1st Qu.: 19224.7   
## Median : 2910 Median : 1.000 Median :53.00 Median : 33170.6   
## Mean : 9757 Mean : 2.088 Mean :53.08 Mean : 39806.9   
## 3rd Qu.: 13168 3rd Qu.: 3.000 3rd Qu.:71.00 3rd Qu.: 53246.2   
## Max. :111850 Max. :10.000 Max. :89.00 Max. :258900.7   
## Outage\_sec\_perweek Tenure MonthlyCharge Bandwidth\_GB\_Year  
## Min. : 0.09975 Min. : 1.000 Min. : 79.98 Min. : 155.5   
## 1st Qu.: 8.01821 1st Qu.: 7.918 1st Qu.:139.98 1st Qu.:1236.5   
## Median :10.01856 Median :35.431 Median :167.48 Median :3279.5   
## Mean :10.00185 Mean :34.526 Mean :172.62 Mean :3392.3   
## 3rd Qu.:11.96949 3rd Qu.:61.480 3rd Qu.:200.73 3rd Qu.:5586.1   
## Max. :21.20723 Max. :71.999 Max. :290.16 Max. :7159.0   
## Churn   
## Length:10000   
## Class :character   
## Mode :character   
##   
##   
##

normalize <- function(x) {  
return ((x - min(x)) / (max(x) - min(x)))  
}  
  
df$Population\_norm<-normalize(df$Population)  
df$Children\_norm<-normalize(df$Children)  
df$Age\_norm<-normalize(df$Age)  
df$Income\_norm<-normalize(df$Income)  
df$Outage\_norm<-normalize(df$Outage\_sec\_perweek)  
df$Tenure\_norm<-normalize(df$Tenure)  
df$MonthlyCharge\_norm<-normalize(df$MonthlyCharge)  
df$Bandwidth\_norm<-normalize(df$Bandwidth\_GB\_Year)  
  
colnames(df)

## [1] "Population" "Children" "Age"   
## [4] "Income" "Outage\_sec\_perweek" "Tenure"   
## [7] "MonthlyCharge" "Bandwidth\_GB\_Year" "Churn"   
## [10] "Population\_norm" "Children\_norm" "Age\_norm"   
## [13] "Income\_norm" "Outage\_norm" "Tenure\_norm"   
## [16] "MonthlyCharge\_norm" "Bandwidth\_norm"

df<-df[,c(9:17)]  
write.csv(df, "c:/users/shua/documents/Data Mining II\_D212/processed\_dataset.csv")

# Part IV: Analysis

To begin the analysis phase the model is built using only two variables at a time. This is because the resulting visualization is difficult for human interpretation if too many variables are used. The k-means model is run multiple times using various variable combination.

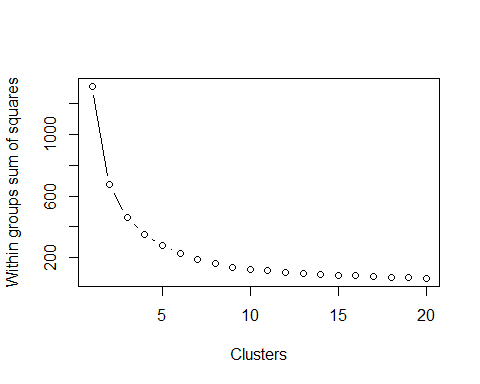
Since the number of clusters needs to be known beforehand, a scree plot is developed to determine the cluster number for each variable combination. A scree plot visualizes the effectiveness of different cluster numbers on the model. A measurement of the model’s effectiveness is the total within cluster sum of squares. Simply put, the model measures the squared distance of each observation to its assigned cluster center and then sums all of squared distances. The lower the number, the more effective the model. By capturing the total within cluster sum of squares for each possible number of clusters (one through twenty), the results are plotted on a scree plot to determine the number of clusters to use for each k-means model.

Having found the number of clusters, this is used as an input to the k-means function for each variable combination. The results are then plotted for analysis of patterns in the data set

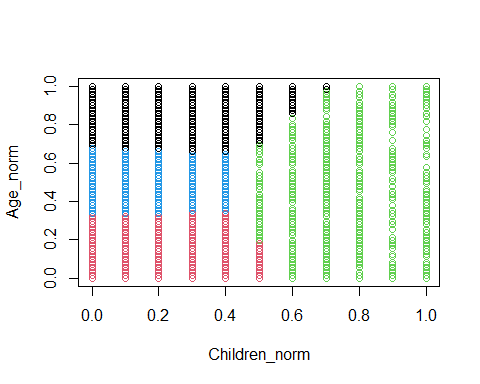
# k-means Tenure & Age  
colnames(df)

## [1] "Churn" "Population\_norm" "Children\_norm"   
## [4] "Age\_norm" "Income\_norm" "Outage\_norm"   
## [7] "Tenure\_norm" "MonthlyCharge\_norm" "Bandwidth\_norm"

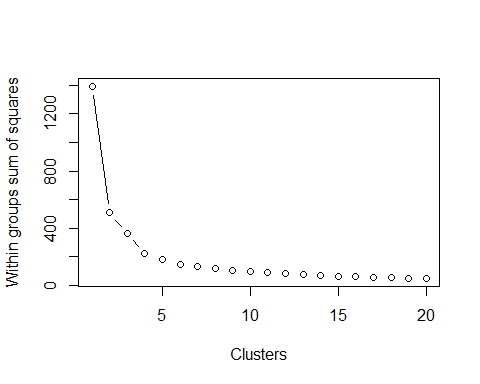
x<-df[,c(3,4)]  
wss<-0  
for (i in 1:20){  
 km.out<-kmeans(x, centers=i, nstart=20, iter.max=30)  
 wss[i]<-km.out$tot.withinss  
}  
  
plot(1:20, wss, type="b", xlab="Clusters", ylab="Within groups sum of squares")



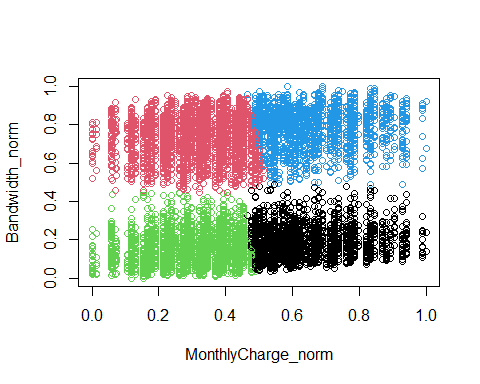
k<-4  
km.out2<-kmeans(x, centers=k, nstart=20)  
plot(x, col=km.out2$cluster, type="p")



#k-means MonthlyCharge & Bandwidth  
x<-df[,c(8,9)]  
wss<-0  
for (i in 1:20){  
 km.out<-kmeans(x, centers=i, nstart=20, iter.max=90)  
 wss[i]<-km.out$tot.withinss  
}  
plot(1:20, wss, type="b", xlab="Clusters", ylab="Within groups sum of squares")



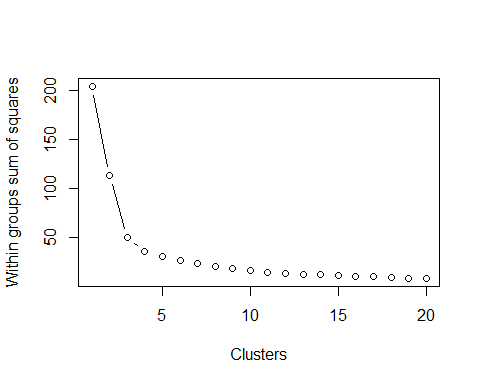
k<-4  
km.out2<-kmeans(x, centers=k, nstart=20)  
plot(x, col=km.out2$cluster, type="p")



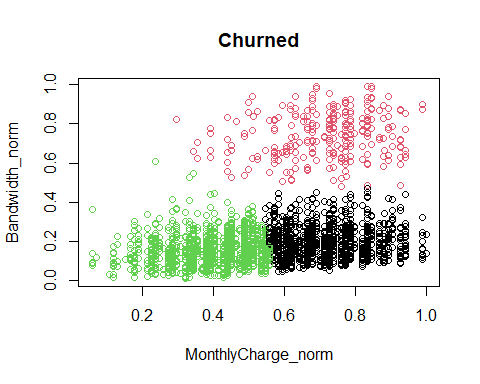
#k-means Churned MonthlyCharge & Bandwidth  
colnames(df)

## [1] "Churn" "Population\_norm" "Children\_norm"   
## [4] "Age\_norm" "Income\_norm" "Outage\_norm"   
## [7] "Tenure\_norm" "MonthlyCharge\_norm" "Bandwidth\_norm"

x<-df[df$Churn=="Yes",c(8,9)]  
wss<-0  
for (i in 1:20){  
 km.out<-kmeans(x, centers=i, nstart=20, iter.max=90)  
 wss[i]<-km.out$tot.withinss  
}  
plot(1:20, wss, type="b", xlab="Clusters", ylab="Within groups sum of squares")



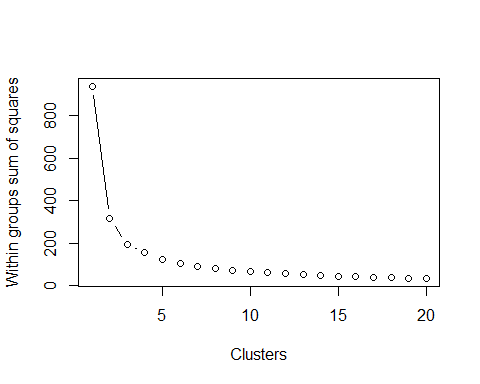
k<-3  
km.out2<-kmeans(x, centers=k, nstart=20)  
plot(x, col=km.out2$cluster, type="p", main="Churned")



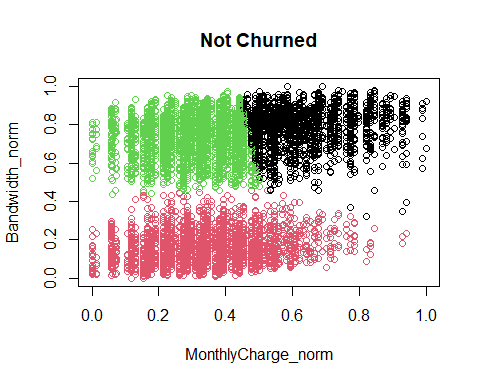
#k-means Not Churned MonthlyCharge & Bandwidth  
colnames(df)

## [1] "Churn" "Population\_norm" "Children\_norm"   
## [4] "Age\_norm" "Income\_norm" "Outage\_norm"   
## [7] "Tenure\_norm" "MonthlyCharge\_norm" "Bandwidth\_norm"

x<-df[df$Churn!="Yes",c(8,9)]  
wss<-0  
for (i in 1:20){  
 km.out<-kmeans(x, centers=i, nstart=20, iter.max=90)  
 wss[i]<-km.out$tot.withinss  
}  
plot(1:20, wss, type="b", xlab="Clusters", ylab="Within groups sum of squares")



k<-3  
km.out2<-kmeans(x, centers=k, nstart=20)  
plot(x, col=km.out2$cluster, type="p", main="Not Churned")

 39 # Part V: Data Summary and Implications

As unsupervised machine learning is a more indefinite exercise than calculating or predicting a particular response, the model accuracy measurements that one would typically use in supervised machine learning are not necessarily applicable in unsupervised machine learning. However, the aforementioned calculation of total within cluster sum of squares is an excellent measurement to determine the efficiency of the k-means model. The lowest total within cluster sum of squares number will result in the more efficient model as this means that the observed data points are closest to the cluster centers.

Most interesting from this analysis is the resulting views from the k-means model using monthly charge and bandwidth used compared between churn and non churn customers. Customers who churned demonstrated a low bandwidth usage frequency generally. There were observations of customers having high bandwidth usage churning but only when the monthly charge was also high. There was not a cluster centered in the quadrant representing high bandwidth usage but low monthly charge. This reflects that high usage customers on affordable plans are mostly retained. When looking at the plot of observations who do not churn impression is inverted. A cluster center is conspicuously absent in the high monthly charge and low bandwidth usage quadrant for non churn customers. This suggest that customers are seeking value. High monthly charge customers are only retained when they also have high bandwidth usage. High bandwidth users who have low monthly charge do not churn.

While these may be intuitive conclusions to guess at without the data analysis, what the k means algorithm also gives us is the boundaries for the the observed clusters. In addition to stating that high monthly charge customers with low bandwidth usage are at risk to churn we can now add that customers having monthly charge in the top 25% and also having bandwidth usage in the lower 50% are at the highest risk to churn. Marketing and customer retention efforts should be directed first and foremost at customers approaching these boundaries. Further, we can now state that customers having monthly charge in the lower 50% and bandwidth usage in the upper 60% are at the lowest risk to churn. Marketing and customer retention energy and resources need not be directed at these customers.

The recommended course of action to the organization is to implement this model for regular use to identify customers at risk to churn. THe boundaries identified in this model can be used as in a targeted and scaled approach to customer retention depending on the customers latest location in the monthly charge and bandwidth usage plot. Doing so should result in higher retention rates and consequently saved costs associated with maintaining the organization’s servicing portfolio.

This analysis had a number of limitations. First, the data set provided is a test data set. Therefore, it may not be reflective of more likely real world scenarios. Part of the analysis endeavor is to be able to test informed assumptions about the data. Working with a test data set, however, may create isolation from real world cause and effect relationships whereby analyst intuition can be neutralized. Secondly, the analysis was limited by the variables provided. Given the opportunity to collect more data points in terms of variables or even observations, such as number or frequency of late payments, this analysis could be expanded. Lastly the analysis is limited by the assumption of the k means algorithm that the number of clusters be known in advance. Especially when working in isolation with test data sets and no real world contacts, the analyst cannot contact a subject matter expert to gather some form of advanced knowledge for the number of clusters or direction of the analysis. Nevertheless, this results of this analysis under the given conditions would carry sufficient validity for consideration of its findings.